ROOT Users Workshop

Contribution ID: **44**

Type: **Presentation**

Distributed RDataFrame: supported backends, latest improvements and future plans

Monday, May 9, 2022 10:55 AM (20 minutes)

The declarative programming model offered by RDataFrame provides high-level abstractions for users to operate on their datasets in a much more ergonomic fashion compared to previous imperative interfaces. This tool has already seen a lot of usage in real-world analyses and production environments, showing optimal results. RDataFrame has always been oriented towards parallelisation, with native support for multi-threading execution on a single machine which doesn't need changes in user code. The parallelisation capabilities have more recently been extended with a Python layer that is capable of steering and executing the RDataFrame computation graph over a set of distributed nodes, also in this case requiring minimal code changes. This new extension features a modular design, such that it can support multiple backends in order to exploit the vast ecosystem of distributed computing frameworks with Python bindings. This talk presents the design behing distributed RDataFrame, discussing the currently available execution backends, the latest improvements and how the tool will continuously evolve in the near future.

Summary

Primary authors: PADULANO, Vincenzo Eduardo (CERN); TEJEDOR SAAVEDRA, Enric (CERN); GUIRAUD, Enrico (EP-SFT, CERN)

Presenter: PADULANO, Vincenzo Eduardo (CERN)

Session Classification: First Session